Predicting the COVID-19 Cases: Combining Standard Methods with Mobility Data

Zichang Ye (zy1545), Duo Jiang (dj1057)
Jiarui Tang (jt3869), Chutang Luo (cl5293), Yutong Chen (yc4127)

Center for Data Science, New York University

Thursday 21$^{\text{st}}$ May, 2020

1 Introduction

1.1 Description of Problem

Coronavirus disease 2019 (abbreviated as COVID-19) outbreak is a global pandemic that started at the end of 2019 and continues through 2020, influencing the world and every person's life unprecedentedly. The United States has declared national emergency on March 13th, followed by travel-restrictions, social distancing measures and close-down of most states. Up until May 17th, there had been 1,500,148 confirmed cases in the US and 348,232 confirmed cases in New York, nearly $\frac{1}{4}$ of the total nationwide confirmed cases. As NYU students, the question of how the pandemic is going to develop and when it's going to end has been our greatest concern lately.

1.2 Approach

In this work, we aim to predict the cumulative confirmed cases in New York State by county. We obtained data of daily confirmed and death cases from onepoint3acres and the Center for Systems and Engineering at JHU as well as social mobility data from SafeGraph COVID-19 Data Consortium and hospital, population data from census bereau. Our approach includes feature engineering of county-wise import risk based on concepts of epidemiology, together with three modeling approaches combining machine learning approaches and epidemiology models adopted from related works.

1.3 Summary of results/conclusion

Our results suggest varying performance for different counties under each modeling framework, in most of the cases underestimating the confirmed cases. We propose two conclusions: firstly, our metric of constructing import risk can contribute to improving predictions. Secondly, some frameworks may work better under particular settings

2 Related work

2.1 SIRD model (1)

This work focuses on the analysis and forecast of COVID-19 spreading in China, Italy and France. It uses simple day-lag maps to verify the universality of COVID-19 epidemic in the three countries, suggesting that a mean-field kinetics of epidemic spreading can be captured to represent the general trend irrespective of the specific country(with variation in parameter).The basic approach in this work is a SIRD model to fit data and analyse the recovery rate, infection rate and death rate. In order to represent the impact of the containment measures, the work considered an infection rate r that varies with time. In our work, we adopted the framework of the SIRD model from here, and proceeded to work with the idea of varying parameters.

2.2 ERF error function (2)

This work aims to provide a functional tool to estimate and forecast deaths across locations. The key idea is to parametrize curves that can be fit to data and model parameters with covariates. Firstly, instead of the general sigmoidal function, the work discovered introduced an ERF error function that would provide a better fit to the data. Next, the work uses a link function, which links covariates to parameters in the ERF function. In general, the relationship can be described as:

$$\begin{aligned}\text{parameter} = &\ LinkFunction(\text{covariate} * \text{multiplier} \\ &+ \text{random effect})\end{aligned}$$

In this case, the covariate is defined as a measure of social distancing constructed with social mobility data from Descartes Labs, Safegraph and Google. Also, the work considered prior knowledge specified as simple priors, box constriants and functional priors. Lastly, the optimization problem is solved with the L-BFGS-B algorithm in Scipy. In our work, we adopted the idea of link function and ERF functions from here and experimented with our own covariate and prior settings.

3 Problem Definition and Algorithm

3.1 Task

We used multiple parametric models such as Logistic Growth models and Gaussian CDF functions to model the confirmed cases. For the purpose of defining a general problem, the model output is denoted as Y(t; $\theta^{(t+1)}$), where θ is parameter vectors. Our goal is to predict the $\theta^{(t+1)}$ based on inputs available at time t.

Given a New York county i, the **inputs** are:

- The number of medical resources in county i;
- The number of visits from county j for $j \in \{j|\text{All counties in United States}\}$ at time t;
- The number of confirmed, recovered, and death cases in each visitors' home county j at time t;
- The population and population density in each visitors' home county j;

The **outputs** are the parameters $\theta^{(t+1)}$, the unknown true parameters that best described the shape of the trajectory of the confirmed cases.

In short, our modeling procedure maps the input into the parameters space, and use the parameters to predict the cases. Namely,

$$X_1^{(t)}, ..., X_n^{(t)} \xrightarrow{\text{f}} \theta^{(t+1)} \xrightarrow{\text{SIRD, ERF}} Y^{(\hat{t}+1)}$$

The machine learning method and non-convex optimization are applied to learn an optimal $f(X_1, ..., X_n) = \theta^{(t+1)}$.

3.2 Algorithm

In our modeling process, we fit a logistic growth model as our baseline model. For further improvement, we implemented three approaches. The first approach is a RNN with LSTM unit, the second approach is a modification of SIRD model and the third approach utilized the ERF error function. The detailed models are described below:

3.2.1 Logistic Growth Model

Due to the generality in the process of virus spreading, pre-assumption can be made on the data. The accumulative number of confirmed cases of virus is usually characterized by increasing growth in the beginning period, but a decreasing growth at a later stage, as you get closer to a maximum. This property can be well captured by Logistic Growth Model with formula

$$\text{cases}(t) = \frac{c}{1 + a * e^{-bt}}$$

in which t is variable denoting time, c, a and b are parameters.

3.2.2 Recurrent Neural Network

One natural choice to model time series data is Recurrent Neural Network(RNN). We apply RNN with LSTM unit to model the normalized newly confirmed cases each day. Our speculation is that RNN makes no assumption about the data and thus may have a harder time doing prediction.

3.2.3 SIRD and modified SIRD(mSIRD)

The SIRD model is a mathematical model widely implemented in the field of epidemiology. In a SIRD model, the whole population is divided into susceptible (S), infected (I), recovered (R), dead (D). And the kinetic of the epidemic evolution is described by differential equations:

- $\frac{ds}{dt} = -rSI$ r: infection rate
- $\frac{dI}{dt} = rSI - (a + d)I$
- $\frac{dR}{dt} = aI$ a: recovery rate
- $\frac{dD}{dt} = dI$ d: death rate

One shortcoming of SIRD model is that it fails to incorporate the effects of social mobility and containment measures. The original work used a variant infectious r to solve it. In our case, since we have social mobility data from SafeGraph COVID-19 Data Consortium with which we can calculate a proxy for import risk in each county in New York State, we try to see if the calculated risk can help us determine how the parameters of SIRD model change. We notice that with data for more days, SIRD model always ends up with a better fit for the data. Therefore, we use linear regression which takes parameters and risk calculated using data now to approximate parameters calculated using future data. Note that the linear regression is built from data early on.

3.2.4 ERF error function

We adapted the general framework of (2) in our approach, with two modifications. Firstly, we use import risk as the source of covariate instead of social mobility, and correspondingly we adapted the prediction label from death rate to cumulative confirmed cases. In epidemiology, the import risk is defined as attack rate quantifying transmissibility, which directly ties to the reproductive number of infection. Thus from the aspect of epidemiology, we think that import risk can be a strong covariate linking to prediction of β, the optimal rate of infection. Secondly, we made slight modifications to the hyperparameter settings.

The ERF error function is specified as:

$$\begin{aligned} f(t; \alpha, \beta, p) &= \frac{p}{2} \left(\Psi(\alpha(t - \beta)\right) \\ &= \frac{p}{2} \left(1 + \frac{2}{\sqrt{\pi}} \int_0^{\alpha(t-\beta)} \exp\left(-\tau^2\right) d\tau\right) \end{aligned}$$

Where the three parameters are: level p: controls the maximum asymptotic level that the rate can reach; slope α: the speed of the infection; inflection β: the time at which the rate of change of D is maximal.

In this case, we consider three specific covariate link functions as:

$$\begin{aligned}\alpha^l &= \exp(\mu_\alpha + \mu_\alpha^l)\\ \beta^l &= (\mu_\gamma + \mu_\gamma^l)Covariate^l\\ p^l &= \exp(\mu_p + \mu_p^l)\end{aligned}$$

μ_α, μ_p and μ_γ respectively capture average behavior of parameters μ_α^l, μ_p^l and covariate multiplier across locations, which are considered as fixed effects. μ_α^l, μ_p^l and μ_γ^l are random effects that respectively adjust $\exp(\mu_\alpha)$, $\exp(\mu_p)$ and covariate multiplier to each location.

Since the curves are highly nonlinear, the nonlinear least squares problem is highly nonconvex, and therefore initialization is important. We adjust initial values, priors as well as bounds of fixed and random effects to explore better models.

4 Experimental Evaluation

4.1 Data

There are five components in the data we used. The first is the cumulative counts of confirmed cases, recovered cases, and deaths, covering date from January 22, 2020 to May 15, 2020. The data is obtained from John Hopkins University (3)and 1Point3Acres(4).

The second component is the mobility data released by SafeGraph (5), a company that purchases, anonymizes, and aggregrates the tracking data of mobile devices. The data provides us with the number of visits to each Point of Interest (POI), as well as the counties of origin from each device, from March 1st, 2020 to May 2nd, 2020,in the State of New York. For the privacy purpose, counties with less than five visitors are not included. We normalized the number of visits by the number of active devices in each visitors' home county.

The third component is the number of hospital and beds data in each New York county, which was retrieved from U.S. Department of Homeland Security (6). The original data provides the address, number of staffs and the number of beds in each hospital in the 50 U.S. States, Washington D.C., U.S. territories of Puerto Rico, Guam, American Samoa, Northern Mariana Islands, Palau, and Virgin Islands. We counted the number of hospitals and summed the number of beds by county fipcode.

The fourth component is the estimated population in 2019 in each county, which was estimated by Census Bureau based on the data from the last census(7).

The fifth component is the land area in each county, which was also retrieved from Census Bureau(8). The population density in each county is then calculated by dividing the estimated population in 2019 by the land area.

We have put all codes and dataset in a github repository.[1]

4.2 Methodology

4.2.1 Feature Engineering

We develop a metric that quantifies the import risk of COVID-19 at each time t for a county j, and incorporate this metric into the SIRD and ERF models. For each county i, the import risk at time t is defined as

$$Risk_{t,i} = \sum_j \frac{n_{t-1,j}}{p_j} \cdot (v_{t-7}) \cdot (1 + g_j) \cdot (\rho_j)$$

- j are the home counties of the visitors;
- $n_{t-1,j}$ represents the confirmed cases at county j at time $t-1$;
- p_j represents the population of the county j;
- g_j represents the growth rate of cases over the past 7 days in county j;
- v_{t-7} represents the average of the number of normalized devices visits over the past 7 days;
- ρ_j represents the population density of county j.

We use the fraction $\frac{n_{t-1,j}}{p_j}$ to estimate the infected proportion in the population at county j. Therefore, $\frac{n_{t-1,j}}{p_j} \cdot v_{t-7}$ approximates the proportion of infected people in the visitors. We further multiply the index by the growth rate and population density, as we believe that counties with higher growth rate and denser population are supposed to have higher risk.

4.2.2 Train-Test Split

In our experiment, the training dataset is the data from 2020-03-01 to 2020-04-29 and test dataset is from 2020-04-30 to 2020-05-09.

4.2.3 SIRD with Risk

We fit one SIRD model for each sub-data in a window of 20 days to get the parameters of the SIRD model. Then we fit a linear regression model to learn the parameters from cumulative confirmed cases, death cases and recover cases, normalized number of visitors, and import risk, namely:

$$\begin{aligned}\text{Cases}^{(t+1,i)} &= \text{SIRD}(\theta^{(t+1,i)})\\ \theta^{(t+1,i)} &= f(\text{Cases}^{(t,i)}, \text{Deaths}^{(t,i)}, \text{Visitors}^{(t,i)}, \text{Risk}^{(t,i)})\end{aligned}$$

To predict the new confirmed cases in the next few days, we first use the new features in the next

[1] https://github.com/JiaruiTang/DS1003-Predicting-the-COVID-19-Cases

few days to predict parameters of SIRD model after the few days. Then we use the SIRD model to predict confirmed cases. We use MSE to minimize when fitting and also use MSE to evaluate the performance of our model.

4.2.4 ERF with Risk

The ERF framework enables us to parametrize the parameters using location-specific covariates, Covariates^l. We believe that the risk metric will inform us about the time at which slopes of the curves are peaked, but not the asymptotic level (p) that the curves will eventually reach or the initial slope (α). Therefore, we parametrized β as:

$$\beta^i = (\mu_\gamma + \mu_\gamma^i) \cdot \text{Risk}^i$$

We first fit a joint models with loose constraints on fixed effects and random effects to get an estimation of the parameters. We then fit the models again using these estimations of parameters as the priors. We repeated these steps for different functional form such as expit function and log-Gaussian CDF.

4.3 Results

4.3.1 SIRD and mSIRD

In most cases (21 out of 31 counties), mSIRD gives a better fit as shown in figure 1 for Putnam. However, it can also hurt the performance for counties that do behave differently from most counties. For example, New York City behaves quite differently from other counties in that it has far more cases and thus the modification worsen the performance a lot as shown in figure 2 (Refer to Appendix 6.1).

4.3.2 ERF Functions

The ERF models predict better in 14 counties out of the total of 62 counties in the test set, covering 16% of total population and 46% of total confirmed cases in New York. For counties where the ERF models are better, the MSE is improved by 45%. Specifically, for New York City, the ERF makes a better prediction of the confirmed cases than the baseline, but not for Putnam (See Appendix 1 for visualization).

4.3.3 RNN

In most counties, RNN beats our baseline model, Logistic Growth Model. However, in a few counties like ALLEGANY where there aren't many people infected, Logistic Growth Model does a better job.

4.4 Discussion

4.4.1 Error Analysis

ERF Models. The ERF models systematically underestimates the confirmed cases in many counties. As an extreme example, the models tremendously underestimate the cases in Nassau county by predicting the peaked time much later than expected. There are several possible reasons that drive such errors. First of all, since we parametrized the inflection point, β, with the import risk, how informative the import risk is critical. In New York City, where travel data is more abundant and informative, so the calculated import risk may be helpful by reducing the bias of the model. In counties such as the Nassau counties, travel data may not be that reliable, and thus can add noise to the model. Secondly, since the ERF framework estimates a joint non-linear models across locations, the hypothesis space is non-convex, initialization is critical. As a result, we may not have found a suitable initialization that works well for all locations.

mSIRD. We are trying to find generality of change of parameters among all counties and thus we can use this information to predict how the risk affects the parameters change. As a consequence, counties that are very difference from other counties suffer from bias. For example, mSIRD for New York City does not work well.

4.4.2 Future Works

- **A More Careful Data Preprocessing.** [2] uses and "cross-validates" four different sources of mobility data to estimate the mobility across regions.
- **Exploration of Initializations.** We can keep exploring different sets of initializations and priors for random effects for optimizing the ERF models.
- **mSIRD.** There is another way to predict parameter changes from risk that is more "time series". That is to use a model like RNN to find how the risk affects the parameters for one county at a time using only this county's history data.

5 Conclusion

By using machine learning methods and GLM framework, we combine current methods of predicting infected cases in an epidemic with new data sources. In particular, our contributions to the academic discussion in three ways. Firstly, we propose and experiment a new algorithm to learn the parameters of SIRD using location-specific information such as confirmed cases and import risk. Secondly, we design a metric to quantify the import risk based on travel data. Thirdly, our experiment show that the information in mobility data can be helpful in predicting the future cases in both the SIRD and GLM frameworks.

6 Appendix

6.1 Figures

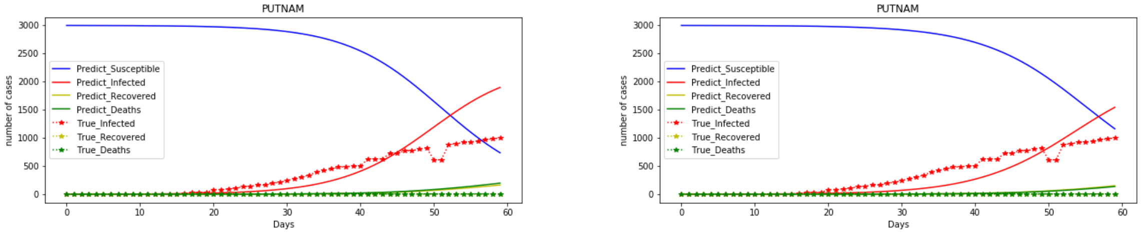

Figure 1: Left: SIRD Plot fitting using data from the first 50 days in Putnam. Right: mSIRD plot using parameters predicted from parameters using in SIRD plot and risk in the fiftieth day in Putnam

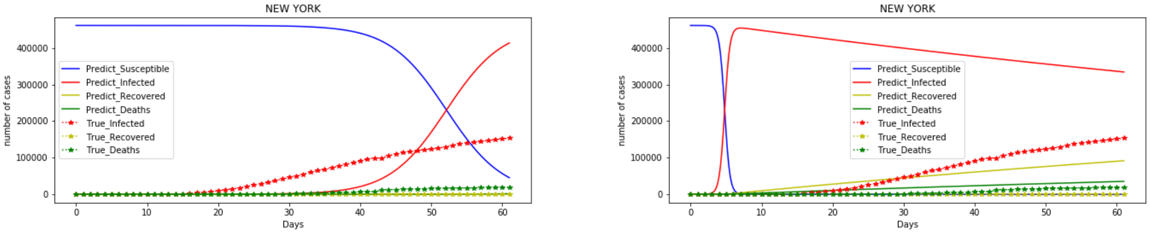

Figure 2: Left: SIRD Plot fitting using data from the first 50 days in New York City. Right: mSIRD plot using parameters predicted from parameters using in SIRD plot and risk in the fiftieth day in New York City

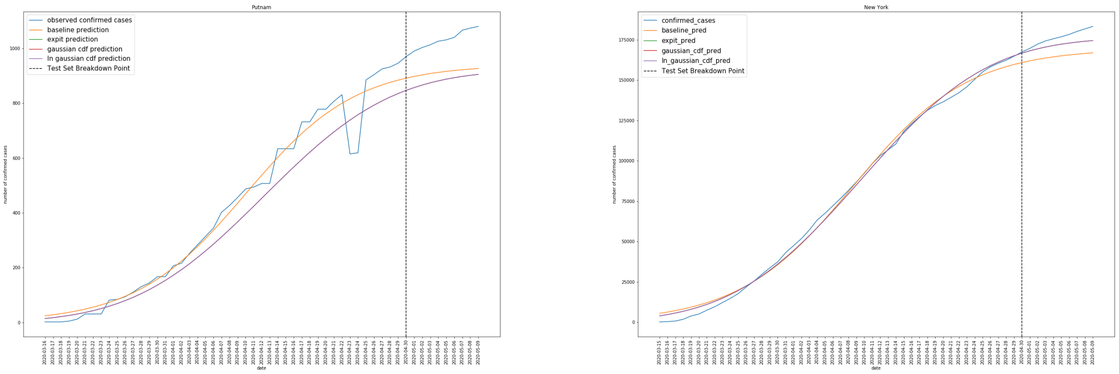

Figure 3: Left: SERF error function Plot fitting using all the data in Putnam. Right: ERF error function Plot fitting using all the data in New York City

6.2 Visualization of social mobility data

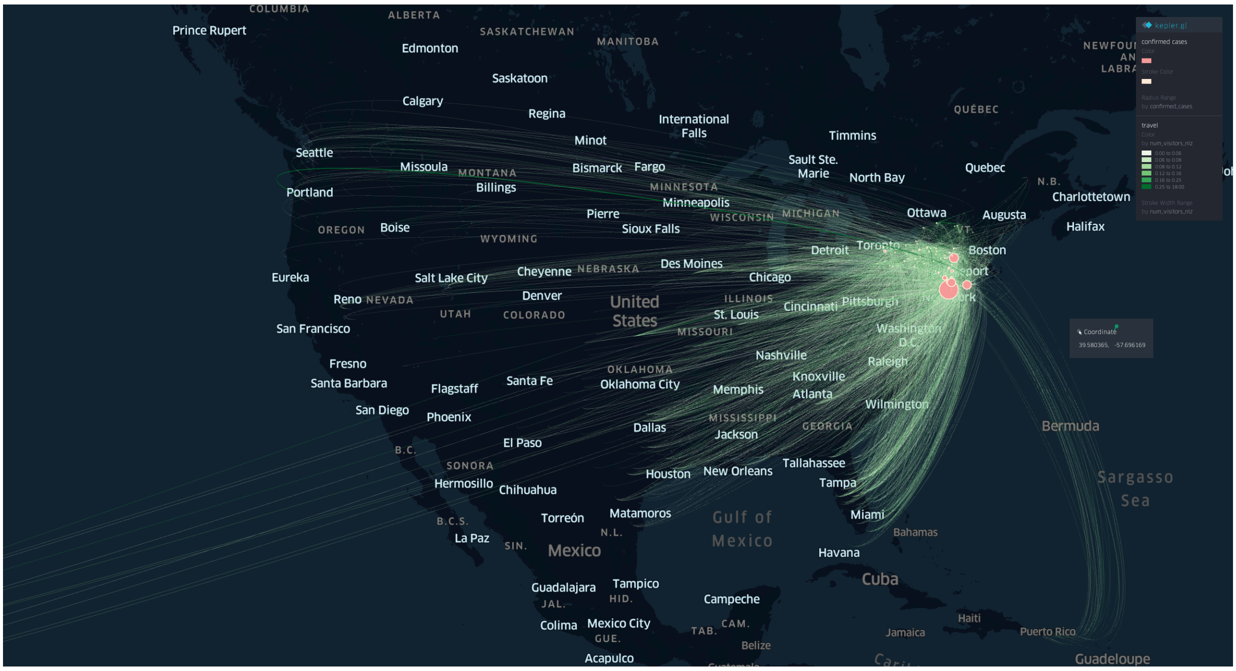

Figure 4: The Travel Patterns Data with New York as Destinations

6.3 Datasets Example

Field name	Description	Example value
FIPS	unique fips code for counties	36001
date	date	2020-03-15
confirmed_cases	cumulative confirmed cases	7
confirmed_cases_smooth	confirmed_cases + 1 to avoid 0 division	8
num_visits_movavg_14	average number of visits in the past 14 days	825.086957
num_visitors_nmlz_movavg_14	normalized moving average	8.216402
risk	calculated imported risk	0.001308
risk_density_weighted	imported risk weighted by population density	0.008328
risk_moving_avg	import risk calculated with moving averaged num_visitor	0.001308
risk_moving_avg_density_weighted	import risk with ma weighted by population density	0.008328
COUNTY	COUNTY name	ALBANY
STATE	state abbrev	NY
num_hospital	number of hospital in the county	7.0
num_beds	number of bed in the county	-617.0
cum_deaths_count	cumulative death	0.0
cum_recovered_count	cumulative recovered	0.0
Population Density	Population density from bereau data	584.364958

6.4 Contribution of Team members

All the members contributed to discussion and final write-up.
Duo Jiang: RNN and SIRD modeling
Zichang Ye: ERF modeling, data preprocessing
Jiarui Tang: Data preprocessing
Chutang Luo: ERF modeling
Yutong Chen: Literature review on modeling

References

[1] D. Fanelli and F. Piazza, “Analysis and forecast of covid-19 spreading in china, italy and france,” 2020.

[2] I. C.-. health service utilization forecasting team and C. J. Murray, “Forecasting the impact of the first wave of the covid-19 pandemic on hospital demand and deaths for the usa and european economic area countries,” 2020.

[3] JHU, “Cssegisanddata.”

[4] 1Point3Acres, “Global covid-19 tracker and interactive charts.”

[5] Safegraph, “Weekly patterns.”

[6] hifld, “hospitals.”

[7] CENSUS, “2010s-counties-total.”

[8] CENSUS, “usa-counties-2011.”

7 Github repository

`https://github.com/JiaruiTang/DS1003-Predicting-the-COVID-19-Cases`